Some New Performance Criteria in Robust Statistics-Small Sample Robustness

by

P. Papantoni-Kazakos

January 1977

Technical Report # 7701

SOME NEW PERFORMANCE CRITERIA IN ROBUST STATISTICS -- SMALL SAMPLE ROBUSTNESS*

by

P. Papantoni-Kazakos
Department of Electrical Engineering
Rice University

ABSTRACT

In this paper the analysis of estimates operating on dependent data is considered. Special dependent data structures are considered and the analysis is made for three different choices of contamination and performance distance measures.

For contamination and performance measures both being Lèvy (Hampel model), an analysis that is particularly oriented toward fast convergence of the estimate to a value that is stable (robust) inside the contaminated family is undertaken. The minimum sample size to satisfy certain performance is investigated and it is found that the problem reduces to finding continuous, absolutely bounded estimates with logarithms of their moment generating function slowly increasing with the absolute value of the argument for all data distributions considered.

1. INTRODUCTION

Humans have always taken satisfaction in the game of "outsmarting Nature." This game becomes particularly intriguing when Nature tries to confuse its human opponents by slightly changing the underlying statistical rules that specify its behavior. To beat Nature's game, then, the human player must follow strategies that are not very sensitive to such statistical changes.

---

* This work was supported by the Air Force Grant AFOSR 77-3156.

The design of such statistical variation-resistant strategies varies with the particular objective of the human player at the time and the decision rules evolving from this design are called robust.

Robust decision rules have been studied by several statisticians [1,3,7, 10,12] and engineers [11,13,18] from different points of view. The promising qualitative analysis of robust estimation first introduced by Hampel [10] has been used and extended [12,18], but the performance criterion has always been expressed through the Lèvy distance of the data as well as the estimator distributions. Specifically, an estimate $\hat{s}(X_n)$ using the data X_n has been called robust (or weakly robust according to [18]) at some distribution Q_o, if every distribution Q that is close to Q_o in the Lèvy distance sense results to $\hat{s}(X_n)$ distribution that is Lèvy distance close to the $\hat{s}(X_n)$ distribution resulting from Q_o-distributed data. This definition of robustness, while leading to constructive analysis of robust estimators [10,18], has two disadvantages: It does not offer convergence rates of the estimates for large number of data and it may be too demanding or even not representative enough for some applications. In addition, the constructive analysis of "Lèvy robust" estimators has been accomplished only for independent X_n components.

In particular, one of the properties that characterize an estimate $\hat{s}(X_n)$ that is "Lèvy robust" at some one-dimensional distribution F_o and is applied on independent data, is continuity at F_o [10,18]. This property, which actually means closeness of the values $\hat{s}(X_n)$, $\hat{s}(X_m)$ for vectors X_n, X_m that specify experimental distributions that are both close to F_o, guarantees convergence of $\hat{s}(X_n)$ to some constant depending on F_o, but it does not specify the convergence rate.

Also, in some applications the preservation of some specific, less general than the Lèvy distance, characteristic may be desirable. In this case the per-

formance criterion is different and so are the desirable properties of the estimator. Performance criteria that are easier to calculate than the Lèvy distance may be preferred.

Finally, the design of estimators that are "robust" in the presence of dependent data structures is certainly a problem that is challenging as well as realistically interesting.

In section 2 of the present paper some preliminary discussion on previous analysis of robustness on statistically contaminated distribution and on possible spaces, different performance criteria is presented.

In section 3, the design of "Vasershtein robust" estimators is undertaken. The dependence structure of the data is naturally incorporated in this case if the distortion or penalty measure is square difference.

In section 4, properties of estimators that are robust as mappings from a data Lèvy-contaminated space to an estimate space characterized by a Vasershtein performance criterion are discussed.

In section 5, a design method of "Lèvy robust estimates" is presented that incorporates an exponential convergence rate. Special dependence structure of the data is considered.

Section 6 includes examples of estimators that are "robust" in the senses of sections 3, 4 and 5.

2. PRELIMINARIES

Robustness has lately been defined [10,18] as stability of some stochastic distance measure defined on the estimator probability space. Specifically, if X_n denotes the vector of n discrete data, Q_o is some well-known multidimensional cumulative distribution applied on X_n (where Q_{on} is the n-dimensional distribution evolving from Q_o), Q is some arbitrary cumulative distribution on X_n; $n = 1,2,\ldots$; $d_1(\cdot,\cdot)$ is a stochastic distance measure

defined on the data distribution space; $\hat{s}_n(X_n)$ is a scalar estimate that is a deterministic function of the data X_n; $D(\hat{s}_n)$, $D^o(\hat{s}_n)$ are the distributions of $\hat{s}_n(X_n)$ determined through Q, Q^o respectively; and $d_{2n}(\cdot,\cdot)$ is a stochastic distance measure defined on the distribution space $D(\hat{s}_n)$, then the sequence $\{\hat{s}_n\}$ is <u>weakly robust</u> at Q^o (as defined in [18]) if given $\epsilon > 0$, there is some $\delta(\epsilon) > 0$ such that: For every Q satisfying $d_1(Q^o,Q) < \delta(\epsilon)$, $d_{2n}(D^o(\hat{s}_n),D(\hat{s}_n)) < \epsilon$; $\forall$ n is implied.

It is obvious from the above definition of weak robustness that the stochastic distances $d_1(\cdot,\cdot)$, $d_{2n}(\cdot,\cdot)$ must be chosen carefully to satisfy the designer's specific objective. In particular, the distance $d_1(\cdot,\cdot)$ represents the kind of Q^o statistical contamination considered, while $d_{2n}(\cdot,\cdot)$ is the performance measure of the estimate.

In the choice of $d_1(\cdot,\cdot)$ and $d_{2n}(\cdot,\cdot)$ the good representation of the particular model as well as the calculability of the distances must be taken into consideration.

In the study presented in [17], it became apparent that among the plethora of stochastic distances existing, there are some more and some less approachable. Specifically, the Lèvy distance although useful in the robust analysis is often very hard to calculate. On the other hand, the Vasershtein distance is simpler in some cases and if also representative of the model considered, it becomes an excellent contamination or performance measure choice.

In the present study, only Lèvy and Vasershtein distance choices will be considered. Therefore, their definition and some of their properties that are related to the analysis in this paper are presented.

Both Lèvy and Vasershtein distances include a distortion or penalty measure $\rho(\cdot,\cdot)$ applied on the outcomes of the distributions involved. Specifically, if we concentrate our attention on discrete data structures, let X_n,

Y_n be two different n-data values and let $\rho(X_n,Y_n)$ be their relative distortion. Also, let Q_n^o, Q_n be two different cumulative distributions of X_n. Then, the Lèvy distance $d_{L_\rho}(Q_n^o,Q_n)$ and the Vasershtein distance $d_{V_\rho}(Q_n^o,Q_n)$ are defined as follows:

$$d_{L_\rho}(Q_n^o,Q_n) = \inf\{\varepsilon: Q_n^o(X_n) \le Q_n(UY_n:\rho(X_n,Y_n)\le\varepsilon)+\varepsilon,$$
$$Q_n(X_n) \le Q_n^o(UY_n:\rho(X_n,Y_n)\le\varepsilon)+\varepsilon;\ \forall X_n\} \quad (1)$$

$$d_{V_\rho}(Q_n^o,Q_n) = \inf E_{Q_{2n}(\cdot,\cdot)}[\rho(X_n,Y_n)] \quad (2)$$
$$\text{all } Q_{2n}(\cdot,\cdot) \text{ inducing}$$
$$Q_n^o,Q_n \text{ marginals}$$

The same distances can be defined for general multivariate distributions Q^o,Q, inducing Q_n^o,Q_n $\forall$ n in the following way:

$$d_{L_\rho}(Q^o,Q) = \sup_n d_{L_\rho}(Q_n^o,Q_n) \quad (3)$$

$$d_{V_\rho}(Q^o,Q) = \sup_n d_{V_\rho}(Q_n^o,Q_n) \quad (4)$$

In [18] it was shown that $d_{L_\rho}(Q_n^o,Q_n)$ is nondecreasing with increasing n for Q_n^o, Q_n being both products of one-dimensional distributions. It is straight forward to show that the same is true for arbitrary Q_n^o, Q_n. The Vasershtein distance $d_{V_\rho}(Q_n^o,Q_n)$, on the other hand, can be independent of n for stationary data structures and proper choice of the distortion (penalty) measure $\rho(\cdot,\cdot)$. Indeed, let $X_n = \{x_i;\ i = 1,\ldots,n\}$ be a sample vector from a wide-stationary process $x(t)$ whose autocovariance function is $R^o(\tau)$ and whose n-dimensional discrete distribution is represented by Q_n^o. Also, let Q_n be the discrete representation of another wide stationary process $y(t)$ with autocovariance function $R(\tau)$. If the Vasershtein distance is defined in the space of jointly stationary distributions and if

$\rho(X_n,Y_n) = \frac{1}{n}[X_n - Y_n]'[X_n - Y_n]$, where $[\ ]'$ means transpose, then as it was found in [15] and [17], one obtains

$$d_{V_\rho}(Q_n^o,Q_n) = \inf_{R^c(0)} \{R^o(0) + R(0) - 2R^c(0) + (m^o - m)^2\} \tag{5}$$

where $R^c(\tau)$ is the crosscovariance function of $x(t)$, $y(t)$, m^o is the mean of $x(t)$ and m the mean of $y(t)$. The expression in (5) is obviously independent of n . Furthermore, if

$$P^o(\lambda) = \sum_{k=-\infty}^{\infty} R^o(k)e^{jk\lambda} \Rightarrow R^o(0) = (2\pi)^{-1}\int_{-\pi}^{\pi} P^o(\lambda)d\lambda \tag{6}$$

$$P(\lambda) = \sum_{k=-\infty}^{\infty} R(k)e^{jk\lambda} \Rightarrow R(0) = (2\pi)^{-1}\int_{-\pi}^{\pi} P(\lambda)d\lambda \tag{7}$$

if Q_n,Q_n^o are Gaussian, then the $R^c(0)$ that satisfies the infimum in (5) was found in [15] to be given by the expression:

$$R^c(0) = (2\pi)^{-1}\int_{-\pi}^{\pi} \sqrt{P^o(\lambda)P(\lambda)}\ d\lambda \tag{8}$$

which provides the following expression of the Vasershtein distance for stationary distribution Gaussian spaces:

$$d_{V_\rho}(Q_n^o,Q_n) = d_{V_\rho}(Q^o,Q) = (2\pi)^{-1}\int_{-\pi}^{\pi} (\sqrt{P^o(\lambda)} - \sqrt{P(\lambda)})^2 d\lambda + (m^o - m)^2 \tag{9}$$

where $\rho(X_n,Y_n) = \frac{1}{n}[X_n - Y_n]'[X_n - Y_n]$ (10)

If Q_n^o, Q_n are not Gaussian, the expression on the left of (9) becomes a lower bound on $d_{V_\rho}(Q_n^o,Q_n)$.

The expression in (9) is simple, computable analytically in most cases, and independent of the sample size n . It is also significant to observe from (9) that the wide sense stationary Gaussian distributions that are "close" with respect to a square error, are the ones whose means and discrete spectra functions are "close."

The constructive analysis of "robust" estimators that was presented in [10] and [18] in addition to being "Lèvy robust," it was also limited in the case of independent data structures. It was found there that the sequence $\{\hat{s}_n(X_n)\}$ of estimates is weakly robust at some Q_{o1} in the sense of both the $d_1(\cdot,\cdot)$, $d_{2n}(\cdot,\cdot)$ distances being Lèvy, if: a) $\hat{s}_n(X_n)$ is continuous for every n as a real function with E^n Euclidean domain, b) $\hat{s}_n(X_n)$ is continuous at Q_{o1}, this continuity meaning that for every X_n, X_m such that they determine experimental distributions $n_{X_n}(y)$, $n_{Y_m}(y)$ that are both Lèvy close to Q_{o1}, it is implied that $|\hat{s}_n(X_n) - \hat{s}_m(Y_m)|$ is small.

The above analysis does not provide convergence rates of the estimate $\hat{s}_n(X_n)$. Such a rate is important to the designer that is looking not only for "robust" estimates but also for sufficient sample sizes also. It is important to have an analysis that answers the double question: "What kind of estimate will be robust for a given contaminated family and how many data are sufficient to guarantee a certain minimum level of performance inside the same family?"

An attempt to answer the above question for certain dependent data structures is even more valuable. Having the dependent data situations in mind, we will first study estimates that are robust for contamination and performance measures that are not both Lèvy. Then, we will present an analysis that although applied to Lèvy distribution contamination and Lèvy performance measure, it defers from the classical one in [10] and [18] in the fact that it can apply to dependent data and it incorporates the convergence rate of the estimates.

3. VASERSHTEIN ROBUST ESTIMATORS

In this section we will assume that the human player has the information that Nature uses a Vasershtein algorithm to contaminate its underline statistics.

That is, we consider the measure of contamination to be the Vasershtein distance and we pick as distortion measure the square error one. If, in addition, Nature picks its statistics from the wide-sense stationary distribution space $\mathcal{F}$ that surrounds a well-known distribution Q^o, then the Vasershtein distance between Q_o and an arbitrary member $Q \in \mathcal{F}$ is given by (as expressed in (9) of section 2):

$$d_{V_\rho}(Q^o,Q) \geq (2\pi)^{-1}\int_{-\pi}^{\pi}(\sqrt{P^o(\lambda)} - \sqrt{P(\lambda)})^2 d\lambda + (m^o - m)^2 \qquad (11)$$

where

$$\rho(X_n,Y_n) = \frac{1}{n}[X_n - Y_n]'[X_n - Y_n] \qquad (12)$$

m^o,m the one-dimensional means corresponding to Q^o and Q and $P^o(\lambda)$, $P(\lambda)$ the respective discrete spectral densities. We have equality in (11) if Q_n^o,Q_n distributions are both Gaussian.

From (11) it is apparent that the Vasershtein ρ-contaminated Gaussian distribution families are the families with contaminated means and spectral densities.

For consideration of arbitrary Q distributions (even when Q^o is Gaussian) we can transform the contamination measure to:

$$d_{P,m}(Q^o,Q) = (2\pi)^{-1}\int_{-\pi}^{\pi}(\sqrt{P^o(\lambda)} - \sqrt{P(\lambda)})^2 d\lambda + (m^o - m)^2 \qquad (13)$$

In other words, we suppose that nature contaminates the data statistics by contaminating the spectral density and the mean of the underline stationary process.

Let us now suppose that the observer's performance measure is the mean square one. That is, if p parameters must be estimated from the collected data X_n, then the robustness of the p-dimensional estimate $\hat{S}_n(X_n)$ is evaluated through a mean square error value. In other words, the estimate designer is fully satisfied if the average mean square distortion between

the $\hat{S}_n(X_n)$ he calculates when Nature uses Q^o underline statistics and the value $\hat{S}_n(Y_n)$ he finds when Q is true, remains small when Q^o, Q are close enough in the Vasershtein distance sense.

If $D^o(\hat{S}_n)$, $D(\hat{S}_n)$ are the distributions of $\hat{S}_n(X_n)$ evolving from Q^o, Q respectively, then

$$d_{V_\rho}(D^o(\hat{S}_n), D(\hat{S}_n)) = p^{-1} \sum_{i=1}^{p} \inf_{r_{ni}(\cdot,\cdot) \text{ inducing } D^o(\hat{s}_{in}), D(\hat{s}_{in})} E_{r_{ni}(\cdot,\cdot)}\{\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)\}^2 \quad (14)$$

where

$$\hat{S}_n(X_n) = \{\hat{s}_{in}(X_n);\ i = 1,\ldots,p\} \quad (15)$$

If

$$E_{D^o(\hat{s}_{in})}\{\hat{s}_{in}(X_n)\} = m^o_{in};\ D_{D(\hat{s}_{in})}\{\hat{s}_{in}(X_n)\} = m_{in} \quad (16)$$

$$E_{D^o(\hat{s}_{in})}\{[\hat{s}_{in}(X_n) - m^o_{in}]^2\} = \sigma^o_{in}(0);\ E_{D(\hat{s}_{in})}\{[\hat{s}_{in}(X_n) - m_{in}]^2\} = \sigma_{in}(0) \quad (17)$$

The following lemma can be expressed:

Lemma 1

The distance in (14) is bounded from below by the expression

$$d_{\sigma,m}(D^o(\hat{S}_n), D(\hat{S}_n)) = \sum_{i=1}^{p} \{(\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)})^2 + (m^o_{in} - m_{in})^2\} \quad (18)$$

Proof

For any joint distribution $r_{ni}(\cdot,\cdot)$ with $D^o(\hat{s}_{in})$, $D(\hat{s}_{in})$ marginals, if $\sigma^i_{in}(0)$ the crosscovariance determined by $r_{ni}(\cdot,\cdot)$, then, the matrix

$$\begin{bmatrix} \sigma^{o}_{in}(0) & \sigma^{c}_{in}(0) \\ \sigma^{c}_{in}(0) & \sigma_{in}(0) \end{bmatrix}$$

must be nonnegative definite.

Therefore,

$$\sqrt{\sigma^{o}_{in}(0)\ \sigma_{in}(0)} \geq \sigma^{c}_{in}(0)$$

Then,

$$E_{r_{ni}(\cdot,\cdot)}\{\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)\}^2 = \sigma^{o}_{in}(0) + \sigma_{in}(0) - 2\sigma^{c}_{in}(0)$$

$$+ (m^{o}_{in} - m_{in})^2 \geq \sigma^{o}_{in}(0) + \sigma_{in}(0) - 2\sqrt{\sigma^{o}_{in}(0)\ \sigma_{in}(0)} + (m^{o}_{in} - m_{in})^2$$

$$= [\sqrt{\sigma^{o}_{in}(0)} - \sqrt{\sigma_{in}(0)}]^2 + (m^{o}_{in} - m_{in})^2$$

and

$$d_{V_\rho}(D^{o}(\hat{S}_n), D(\hat{S}_n)) \geq p^{-1} \sum_{i=1}^{p} \{[\sqrt{\sigma^{o}_{in}(0)} - \sqrt{\sigma_{in}(0)}]^2 + (m^{o}_{in} - m_{in})^2\}$$

The expression in (18) is equal to the distance in (14) if and only if $D(\hat{s}_{in})$ can be the distribution of a linear transformation of the variable distributed as in $D^{o}(\hat{s}_{in})$, for every $1 \leq i \leq p$.

We will summarize the observations we made up to now by the following three definition:

Definition 1

We will call a sequence $\{\hat{S}_n\}$ of p-dimensional estimates ρ-Vasershtein weakly robust inside a wide sense stationary distribution family $\mathcal{F}$ and at some Q^{o} if and only if given $\epsilon > 0$, there is some $\delta(\epsilon) > 0$ such that: For $P^{o}(\lambda), m^{o}$ being the spectral density and the mean induced by Q^{o}, $P(\lambda), m$ being the spectral density and mean of some $Q \in \mathcal{F}$ and for

$$(2\pi)^{-1}\int_{-\pi}^{\pi}(\sqrt{P^o(\lambda)} - \sqrt{P(\lambda)})^2 d\lambda + (m^o - m)^2 < \delta(\epsilon)$$

it is implied that:

$$p^{-1}\sum_{i=1}^{p}\{[\sqrt{\sigma_{in}^o(0)} - \sqrt{\sigma_{in}(0)}]^2 + (m_{in}^o - m_{in})^2\} < \epsilon \ ; \ \forall \ n$$

where $\sigma_{in}^o(0)$, $\sigma_{in}(0)$, m_{in}^o, m_{in} are given by (17) and (16) respectively.

Observations

1. If the $\mathcal{F}$ family in definition 1 is a Gaussian stationary family and the estimates $\hat{s}_{in}(X_n)$ are linear transformations of the data, then the expression

$$(2\pi)^{-1}\int_{-\pi}^{\pi}(\sqrt{P^o(\lambda)} - \sqrt{P(\lambda)})^2 + (m^o - m)^2$$

$$p^{-1}\sum_{i=1}^{p}\{(\sqrt{\sigma_{in}^o(0)} - \sqrt{\sigma_{in}(0)})^2 + (m_{in}^o - m_{in})^2\}$$

are the exact Vasershtein distances of the distributions Q^o, Q and $D^o(\hat{s}_{in})$, $D(\hat{s}_{in})$, respectively.

2. It is evident from definition 1 that since we want the closeness of the estimate means and variances guaranteed by the closeness of just the spectral densities and the means of the data distributions, we must limit the estimates to linear transformations of the observations.

Let us define the linear estimates

$$\hat{s}_{in}(X_n) = \sum_{k=1}^{n} a_{ni}(k)x_k \qquad (19)$$

where $X_n = \{x_k ; k = 1,\ldots,n\}$ and the coefficients $a_{ni}(k)$ are real, scalar, and, in general, different for different dimensionality n. To study robustness, we need the means and variances of the estimate in (19) under data

distributions Q^o and Q. Indeed, we have

$$m^o_{in} = E\{\hat{s}_{in}(X_n)/Q^o \text{ distr}\} = m^o \sum_{k=1}^{n} a_{ni}(k)$$

$$m_{in} = E\{\hat{s}_{in}(X_n)/Q \text{ distr}\} = m \sum_{k=1}^{n} a_{ni}(k)$$

$$\sigma^o_{in}(0) = \sum_{k,\ell=1}^{n} a_{ni}(k)a_{ni}(\ell)E\{(x_k - m^o)(x_\ell - m^o)\} \tag{20}$$

$$= \sum_{k,\ell=1}^{n} a_{ni}(k)a_{ni}(\ell)R^o(k - \ell)$$

$$= (2\pi)^{-1}\int_{-\pi}^{\pi} P^o(\lambda) \sum_{k,\ell=1}^{n} a_{ni}(k)a_{ni}(\ell)e^{-j(k-\ell)\lambda}\, d\lambda$$

where Q^o is wide sense stationary with autocovariance $R^o(\tau)$ and power spectral density $P^o(\lambda)$.

From the above expression we finally obtain:

$$\sigma^o_{in}(0) = (2\pi)^{-1}\int_{-\pi}^{\pi} P^o(\lambda) \left\| \sum_{k=1}^{n} a_{ni}(k)e^{-jk\lambda} \right\|^2 d\lambda \tag{21}$$

and similarly:

$$\sigma_{in}(0) = (2\pi)^{-1}\int_{-\pi}^{\pi} P(\lambda) \left\| \sum_{k=1}^{n} a_{ni}(k)e^{-jk\lambda} \right\|^2 d\lambda \tag{22}$$

for the variance under distribution Q. Applying the Schwartz inequality:

$$\int_{-\pi}^{\pi} \sqrt{P^o(\lambda)}\,\sqrt{P(\lambda)} \left\| \sum_{k=1}^{n} a_{ni}(k)e^{-jk\lambda} \right\|^2 d\lambda \le \left[\int_{-\pi}^{\pi} P_o(\lambda) \left\| \sum_{k=1}^{n} a_{ni}(k)e^{-jk\lambda} \right\|^2 d\lambda\right]^{\frac{1}{2}} \cdot$$

$$\cdot \left[\int_{-\pi}^{\pi} P(\lambda) \left\| \sum_{k=1}^{n} a_{ni}(k)e^{-jk\lambda} \right\|^2 d\lambda\right]^{\frac{1}{2}}$$

on $\left[\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)}\right]^2$

we obtain from (21) and (22):

$$[\sqrt{\sigma^{o}_{in}(0)} - \sqrt{\sigma_{in}(0)}]^2 \leq (2\pi)^{-1}\int_{-\pi}^{\pi}[\sqrt{P^{o}(\lambda)} - \sqrt{P(\lambda)}]^2 \| \sum_{k=1}^{n} a_{n_i}(k)e^{-jk\lambda}\|^2 d\lambda \tag{23}$$

while from (21) we get directly:

$$(m^{o}_{in} - m_{in})^2 = (m^{o} - m)^2[\sum_{k=1}^{n} a_{ni}(k)]^2 \tag{24}$$

The expressions in (23) and (24) express the connection between data and estimate statistics needed to specify ρ-Vasershtein weak robustness, as given by definition 1.

From the analysis done above, and the expression of ρ-Vasershtein weak robustness in definition 1, a lemma offering a constructive properties of estimates that are Vasershtein weak robust is obtained.

Lemma 2

A p-dimensional estimate $\hat{S}_n(X_n)$ that is ρ-Vasershtein weak robust, as expressed by definition 1, must be linear. If this linear estimate is given by the expressions:

$$\hat{S}_n(X_n) = \{\hat{s}_{in}(X_n);\ i = 1,\ldots,p\},\ \hat{s}_{in}(X_n) = \sum_{k=1}^{n} a_{n_i}(k)x_k$$

a sufficient condition for the present sense of robustness is that the sequences

$$\{\sum_{k=1}^{n} |a_{n_i}(k)|\}$$

converge to some finite value for every $1 \leq i \leq p$.

Proof:

Let the sequence $\{\sum_{k=1}^{n} |a_{ni}(k)|\}$ converge to some $A_i < \infty$. Then, also

$$(\sum_{k=1}^{n} a_{n_i}(k))^2 \leq (\sum_{k=1}^{n} |a_{n_i}(k)|)^2 \leq A_i^2 \leq \max_{1\leq i\leq p} A_i^2;\ \forall\, n\ .$$

Given $\epsilon > 0$, pick

$$\delta(\epsilon) = \frac{\epsilon}{2p \cdot \max_{1 \le i \le p} A_i^2}$$

and the conditions in definition 1 are satisfied.

According to lemma 2, the experimental mean $\hat{s}_{in}(X_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$ is a ρ-Vasershtein weakly robust estimate. The condition of lemma 2 can be easily seen to satisfy robustness (not just weak robustness) properties inside some data-distribution contaminated family.

Concluding this section, we want to emphasize that the robustness structure presented here considers dependent data with dependence expressed by arbitrary wide-sense stationary distributions. This dependence was explicitly incorporated in the robustness only through the spectral densities of these distributions.

4. LÈVY-VASERSHTEIN ROBUST ESTIMATORS

In this section we consider the case that the contamination measure on the data distribution space is the Lèvy distance, while the performance measure on the space of the estimates is the Vasershtein-type distance

$$d_{Vt}(D^o(\hat{S}_n), D(\hat{S}_n)) = p^{-1} \sum_{i=1}^{p} \{[\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)}]^2 + (m^o_{in} - m_{in})^2 \quad (25)$$

The characteristics $\sigma^o_{in}(0)$, $\sigma_{in}(0)$, m^o_{in}, m_{in} are given by expressions (16) and (17) of the previous section and the distance in (25) is equal to the Vasershtein distance again if the data distribution family is a wide sense stationary family and the estimates are linear.

The robustness considered here is precisely expressed by the following definition.

Definition 2

A sequence of p-dimensional estimates $\{\hat{S}_n(X_n)\}$ is weakly ρ-Lèvy-Vasershtein robust at some distribution Q^o if and only if: given $\epsilon > 0$, there is some $\delta(\epsilon) > 0$ such that for every distribution Q satisfying:

$$d_{L_\rho}(Q^o,Q) < \delta(\epsilon)$$

it is implied that

$$d_{V_t}(D^o(\hat{S}_n),D(\hat{S}_n)) < \epsilon \; ; \; \forall \, n \; .$$

The distance $d_{L_\rho}(Q^o,Q)$ is defined by expressions (1) an (3) and the distance $d_{V_t}(D^o(\hat{S}_n),D(\hat{S}_n))$ by expression (25). The dependence structure of the data (as expressed by Q^o and Q) is arbitrary at this point.

In the analysis for the discovery of constructive properties of the estimates that are robust in the ρ-Lèvy-Vasershtein sense, we will need to bound the absolute values of the $\hat{S}_n(X_n)$ components $\hat{s}_{in}(X_n)$; $i = 1,\ldots,p$, for every i and every n. That restriction is mostly very useful realistically whenever we are seeking the estimation of parameters whose values (we know in advance) move inside a limited interval. The value restriction on the estimates rejects a priori the unacceptably (or dangerously) false decisions on the parameters of interest.

Similarly to the method presented in [18], we will brake the constructive analysis of the ρ-Lèvy-Vasershtein weakly robust estimates into two parts. One for sample sizes n bounded from above by some n_o and one for the n's that exceed this bound n_o.

We proceed first with the bounded n part, presenting the following lemma.

Lemma 3

Let an estimate $\hat{S}_n(X_n) = \{\hat{s}_{in}(X_n);\ 1 \leq i \leq p\}$ be absolutely bounded for

every i, n, X_n and let all the $\hat{s}_{in}(X_n)$ components be continuous as real functions defined on the E^n Euclidean space $\forall\, n$. Then, given some natural number n_o and some $\epsilon > 0$, there is some $\delta(\epsilon, n_o) > 0$ such that $\forall\, n \leq n_o$ and for Q_n:

$$d_{L_\rho}(Q_n^o, Q_n) < \delta(\epsilon)$$

it is implied that:

$$d_{V_t}(D^o(\hat{S}_n), D(\hat{S}_n)) < \epsilon$$

The proof of this lemma is presented in appendix A. The continuity of $\hat{s}_{in}(X_n)$ as a real function is from the $\rho(X_n, Y_n)$ distortion measure on the data that is incorporated in the Lèvy distance, to the absolute value difference $|\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)|$ of the estimates. No consideration of particular dependence structure of the data was necessary at this point.

The lemma we present next combines the properties of the estimator that satisfy the ρ-Lèvy-Vasershtein weak robust requirements for finite as well as infinite sample sizes n.

For the transition to the infinite n step, the specification of a particular dependence structure of the data is necessary. In particular, we will assume that the family of data statistics considered is limited to m-dependent distributions. In addition, the data will be collected in groups of k consecutive data and the groups will be in distance of m data from each other. Specifically, the data vector X_n will consist, in this case, of k-dimensional vectors X_{ki}; $i = 1, 2, \ldots, n_1$. The components of each X_{ki} vector are dependent, but X_{ki} is independent of X_{kj} for $i \neq j$.

The experimental distribution of the vector X_n is defined then as follows

$$n^k_{X_n}(\omega_1,\ldots,\omega_k) = \frac{1}{n_1}\{\#\text{ of } X_{ki}\text{'s with } x_1 < \omega_1,\ x_2 < \omega_2,\ldots,x_k < \omega_k\} \tag{26}$$

A continuity of the estimate that is actually a property of the estimator indicating stochastic stability around the data central distribution Q^o is defined below:

Definition 3

For data X_n consisting of X_{ki} independent vectors and experimental distributions defined by (26), an estimator $\hat{S}_n(X_n) = \{\hat{s}_{in}(X_n);\ 1 \le i \le p\}$ is continuous at Q^o_k if and only if:

Given $\epsilon > 0$, there is some $\mu(\epsilon) > 0$, n_o such that: For every X_n, Y_n satisfying

$$d_{L_\rho}(n^k_{X_n},\ Q^o_k) < \mu(\epsilon)$$

$$d_{L_\rho}(n^k_{Y_m},\ Q^o_k) < \mu(\epsilon)$$

and $n > n_o$

it is implied:

$$\max_i |\hat{s}_{in}(X_n) - \hat{s}_{im}(Y_m)| < \epsilon$$

Combination of lemma 3 and definition 3 leads to the following lemma whose proof can be found in Appendix A.

Lemma 4

Let $\hat{S}_n(X_n) = \{\hat{s}_{in}(X_n);\ 1 \le i \le p\}$ be absolutely bounded for every i, n, X_n. Let every $\hat{s}_{in}(X_n)$ be a ρ-continuous real function on E^n $\forall$ n and let $\hat{S}_n(X_n)$ be continuous at Q^o_k where X_n is formed from n_1 independent k-dimensional data vectors. Then the estimate $\hat{S}_n(X_n)$ is ρ-Lèvy-Vasershtein weakly robust.

The ρ included in lemma 4 and definition 2 indicates the distortion measure included in the Lèvy distance while, as we show in the previous section, the Vasershtein type performance $d_{V_t}(\cdot,\cdot)$ is the result of the consideration of the square error data measure in the expression of the Vasershtein distance.

The estimator properties that guarantee weak robustness through the contamination and performance measures considered in this section are similar to the ones of the classical Lèvy-Lèvy model analyzed in [10] and [18]. Here the absolute boundness of the estimate is an additional desirable property.

The estimators that are not weakly robust in the Lèvy-Lèvy sense are not robust in the Lèvy-Vasershtein sense also.

The means for the design of the properly "robust estimates are similar in both of the above cases.

In the following section we present an alternative analysis method that incorporates convergence rates and gives us a better feeling as to the proper design methods for Lèvy contamination, Lèvy performance robust estimates.

5. A NEW APPROACH TO LÈVY-LÈVY ROBUST ESTIMATION

The robust models that were considered by Hampel [10] and Papantoni-Kazakos [18] were based on Lèvy-contaminated data distribution families and Lèvy performance criterion of the estimates. Specifically, if $\rho(X_n,Y_n)$ is some distortion (penalty) measure defined on the data, robustness is defined as follows according to this model:

Definition 4

A sequence $\{\hat{\S}_n(X_n)\}$ of p-dimensional estimates is Lèvy-Lèvy weakly robust at Q^o if and only if given $\epsilon > 0$, there is some $\delta(\epsilon) > 0$ such that:

For every Q such that

$$d_{L_\rho}(Q^o,Q) < \delta(\epsilon)$$

it is implied that:

$$d_{L_\rho}(D^o(\hat{S}_n),D(\hat{S}_n)) < \epsilon \ ; \ \forall \ n$$

$D^o(\hat{S}_n)$, $D(\hat{S}_n)$ are the estimate p-dimensional distributions induced by $Q^o(X_n)$, $Q(X_n)$, respectively. Also, the distributions Q^o, Q generate, in general, dependent X_n vectors.

If n_o is some finite natural number, the stability property expressed by definition 4 is satisfied for $n \leq n_o$ if the components $\hat{s}_{in}(X_n)$ of the estimate $\hat{S}_n(X_n)$ are all continuous as functions with E^n Euclidean domain, where n any natural number. The proof of this is similar to the corresponding proof for independent data appearing in [18] and to the proof of lemma 3 of the present paper that can be found in appendix A. Formally speaking:

Lemma 4

If $\hat{S}_n(X_n)$ is continuous as a function on E^n $\forall$ n , then, given $\epsilon > 0$ and n_o , there is some $\delta(\epsilon,n_o) > 0$ such that:

For every distribution Q satisfying

$$d_{L_\rho}(Q^o,Q) < \delta(\epsilon,n_o)$$

it is implied:

$$d_{L_\rho}(D^o(\hat{S}_n),D(\hat{S}_n)) < \epsilon \ ; \ \forall \ n \leq n_o$$

For data samples that are unlimited in number, we would like to investigate properties of the estimates that, in addition to satisfying the conditions in definition 4, they also guarantee fast convergence. Such an analysis will provide the designer with the additional valuable information of the sample sizes necessary to satisfy a given performance. The performance measure in

this case is, of course, the Lèvy deviation of the estimate whenever the data distributions move inside a certain sphere.

Before we proceed in the analysis, we need the assumption of a certain dependence structure of the data. As in the previous section, we will assume that the data vector X_n consists of n_1 k-dimensional independent vectors X_{kj} ; $j = 1,\ldots,n_1$; $n = kn_1$.

To avoid unnecessary generalities and to make the analysis more meaningful, we will also consider a particular, quite general form of estimates. Specifically, we will assume that to estimate the component s_i of the parameter vector S_p we apply a different, in general, continuous transformation on each of the first q k-size data blocks and form a linear combination of these transformations. We repeat the same transformation to the next qk data block, etc. and we finally average out the resulting values. The assumption is, of course, that we always receive data in qk size blocks. To express the above description mathematically, we write:

$$\hat{s}_i(X_{nqk}) = \frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{q} a_\ell \mu_{i\ell}(X_{k,j\ell}) \tag{27}$$

where $X_{k,j\ell}$ the $j\ell$th k-size block of data from the vector X_{nqk} , and $\mu_{i\ell}(\cdot)$ a continuous scalar function on the E^k Euclidean space for every $1 \leq \ell \leq q$. The continuity of $\mu_{i\ell}(\cdot)$ guarantees satisfaction of lemma 4. We consider estimators of the same general nature for all the S_p components, therefore we finally obtain a system of estimates as given by (27) for $1 \leq i \leq p$. We will observe at this point that since the k-dimensional data vectors X_{kj} have been assumed to be independent from each other, the functions $\mu_{i\ell}(X_{k,j\ell})$ are independent random variables for different $j\ell$'s.

For convenience, we will pick here the distortion measure ρ that is included in the Lèvy distances of definition 4 to be given by:

$$\rho_o(X_n,Y_n) = \frac{1}{n}\sum_{i=1}^{n} |x_i - y_i|;\ X_n = \{x_i;\ i = 1,\ldots,n\};$$

$$Y_n = \{y_i;\ i = 1,\ldots,n\} \tag{28}$$

Some different distortion (penalty) measures $\rho(\cdot,\cdot)$ lead to analysis similar to the one that will be presented in this section, therefore the choice in (28) is not truly restrictive.

Starting on the analysis of the estimators in (27), let us suppose that for some $\epsilon > 0$, some n, and some Q^o, Q^1 distributions we have:

$$d_{L_{\rho_o}}(D^o(\hat{S}_{nqk}),D^1(\hat{S}_{nqk})) < \epsilon \tag{29}$$

According to Strassen ([2], Th. 11) the condition (29) is true if and only if there is some 2p-dimensional distribution $D^{o1}(\cdot,\cdot)$ with $D^o(\cdot)$, $D^1(\cdot)$ marginals such that

$$D^{o1}\{\frac{1}{p}\sum_{i=1}^{p}\frac{1}{n}\Big|\sum_{j=1}^{n}\sum_{\ell=1}^{q} a_\ell[\mu_{i\ell}(X_{k,j\ell}) - \mu_{i\ell}(Y_{k,j\ell})]\Big| \geq \epsilon\} \leq \epsilon \tag{30}$$

In (30), the estimate form in (27) is considered and $X_{k,j\ell}$, $Y_{k,j\ell}$ are distributed as in Q^o, Q^1 respectively.

The joint distribution D^{o1} can be translated to a distribution Q^{o1} with Q^o, Q^1 marginals instead through a specific estimate choice. Since the Q^o, Q^1 distributions are representing data of k-size independent blocks, Q^{o1} and D^{o1} will be such that they maintain this independence. In other words, D^{o1} in (30) should be such that the differences $[\mu_{i\ell}(X_{k,j\ell}) - \mu_{i\ell}(Y_{k,j\ell})]$ are independent from each other for different $j\ell$ values.

Observing expression (30), we see that due to the truth of the inequality:

$$\sum_{i=1}^{p} D^{o1}\{\frac{1}{n}\Big|\sum_{j=1}^{n}\sum_{\ell=1}^{q} a_\ell[\mu_{i\ell}(X_{k,j\ell}) - \mu_{i\ell}(Y_{k,j\ell})]\Big| \geq \epsilon\} \geq$$

$$\geq D^{o1}\{\bigcup_{1\leq i\leq p}[\frac{1}{n}\,|\sum_{j=1}^{n}\sum_{\ell=1}^{q}a_\ell[\mu_{i\ell}(X_{k,j\ell})-\mu_{i\ell}(Y_{k,j\ell})]\,|\geq\epsilon]\}$$

$$\geq D^{o1}\{\frac{1}{p}\sum_{i=1}^{p}\frac{1}{n}|\sum_{j=1}^{n}\sum_{\ell=1}^{q}a_\ell[\mu_{i\ell}(X_{k,j\ell})-\mu_{i\ell}(Y_{k,j\ell})]\,|\geq\epsilon\}$$

to have (29) satisfied, it is sufficient that:

$$D^{o1}\{\frac{1}{n}\,|\sum_{j}^{n}\sum_{\ell=1}^{q}a_\ell[\mu_{i\ell}(X_{k,j\ell})-\mu_{i\ell}(Y_{k,j\ell})]\,|\geq\epsilon\}\leq\frac{\epsilon}{p}\;;\;\forall\; i{:}1\leq i\leq p \tag{31}$$

So, if for some D^{o1} choice with Q^o, Q marginals, (31) is satisfied, so is (29). For additional simplification we will assume that we are working on distribution spaces that generate stationary data. Then,

$$Q^o(X_{k,j\ell}=Z_k)=Q^o(X_{k,mr}=Z_k)\;;\;\forall\; j\ell,\ mr$$

$$Q^1(Y_{k,j\ell}=Z_k)=Q^1(Y_{k,mr}=Z_k)\;;\;\forall\; j\ell,\ mr$$

In this case the distances $d_{L_{p_o}}(Q^o,Q)$, $d_L(D^o(\hat{S}_n),D(\hat{S}_n))$ are generated by the k-dimensional distributions Q_k^o, Q_k^1. Also, we can then define:

$$m_{i\ell}^o=E_{Q^o}\{\mu_{i\ell}(X_{k,j\ell})\}$$
$$m_{i\ell}^1=E_{Q^1}\{\mu_{i\ell}(Y_{k,j\ell})\} \tag{32}$$

where $m_{i\ell}^o$, $m_{i\ell}^1$ are independent of $j\ell$. From (32) and (31) we obtain that if we want (29) satisfied, it is sufficient to require the following condition:

$$D^{o1}\{|\frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{q}a_\ell[[\mu_{i\ell}(X_{k,j\ell})-m_{i\ell}^o]-[\mu_{i\ell}(Y_{k,j\ell})-m_{i\ell}^1]]+$$
$$+\sum_{\ell=1}^{q}a_\ell[m_{i\ell}^o-m_{i\ell}^1]\,|\geq\epsilon\}\leq\frac{\epsilon}{p}\;;\;\forall\; i{:}1\leq i\leq p \tag{33}$$

where

$$E_{D^{o1}}\{\mu_{i\ell}(X_{k,j\ell})-\mu_{i\ell}(Y_{k,j\ell})\}=m_{i\ell}^o-m_{i\ell}^1 \tag{34}$$

Directly from (33) we can express the following stronger condition, to guarantee

satisfaction of (29):

$$D^{ol}\{\left|\frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{q}a_\ell[[\mu_{i\ell}(X_{k,j\ell}) - m^o_{i\ell}] - [\mu_{i\ell}(Y_{k,j\ell}) - m^1_{i\ell}]]\right|$$

$$\geq \epsilon - \sum_{\ell=1}^{q}|a_\ell||m^o_{i\ell} - m^1_{i\ell}|\} \leq \frac{\epsilon}{p}\ ;\ \forall\ i{:}1 \leq i \leq p \qquad (35)$$

At this point we find necessary to summarize our analysis up to this point in the following Corollary:

Corollary 1

If, given $\epsilon > 0$, there is some n_a and some $\delta(\epsilon,n_a) > 0$ such that: $\forall\ n \geq n_a$ and every Q^1_k satisfying:

$$d_{L_{\rho_o}}(Q^o_k,Q^1_k) < \delta(\epsilon,n_a)$$

there is some D^{ol}_k with marginals Q^o_k, Q^1_k implying:

$$D^{ol}_k\{\left|\frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{q}a_\ell[[\mu_{i\ell}(X_{k,j\ell}) - m^o_{i\ell}] - [\mu_{i\ell}(Y_{k,j\ell}) - m^1_{i\ell}]]\right|$$

$$\geq \epsilon - \sum_{\ell=1}^{q}|a_\ell||m^o_{i\ell} - m^1_{i\ell}|\} \leq \frac{\epsilon}{p}\ ;\ \forall\ 1 \leq i \leq p \qquad (36)$$

where $\mu_{i\ell}(\cdot)$ continuous on E^k for $1 \leq \ell \leq q;\ 1 \leq i \leq p$, then the estimate described by (27) is Lèvy-Lèvy weakly robust according to definition 4.

Now that we have summarized the observations up to now, we will continue with analysis of the conditions of corollary 1.

Observing condition (36) we realize that if we want it satisfied for given $\epsilon > 0$, the sum

$$\sum_{\ell=1}^{q}|a_\ell||m^o_{i\ell} - m^1_{i\ell}| \qquad (37)$$

must be smaller than ϵ . In other words, for given $\epsilon > 0$, n_a we would like to find some $\delta(\epsilon,n_a) > 0$ that first guarantees the satisfaction of the inequality

$$\sum_{\ell=1}^{q} |a_\ell| |m_{i\ell}^o - m_{i\ell}^1| < \epsilon_1 \text{ ; where } \epsilon_1 > 0 \text{ and such that } \epsilon_1 < \epsilon \tag{38}$$

for every $Q_k^1 : d_{L,\rho_o}(Q_k^o, Q_k^1) < \delta(\epsilon,n_a)$ and then it secures the existence of some D_k^{o1} with Q_k^o, Q_k^1 marginals that satisfies (36).

Let us pick $\epsilon_1 = \epsilon/2$. Then, if $\mu_{i\ell}(X_k)$ is bounded absolutely by some constant B for every $1 \leq i \leq p$, $1 \leq \ell \leq q$ and X_k , we know from lemma 3 and its proof in appendix A that given

$$\frac{\epsilon}{2 \sum_{\ell=1}^{q} |a_\ell|}$$

there is some

$$\delta\left(\frac{\epsilon}{2 \sum_{\ell=1}^{q} |a_\ell|}, k, B\right) > 0$$

such that

$$\sum_{\ell=1}^{q} |a_\ell| |m_{i\ell}^o - m_{i\ell}^1| < \frac{\epsilon}{2}$$

After this last observation, we can go one step further and express the following corollary that is a simplified form of corollary 1:

Corollary 2

Let the set of $\mu_{i\ell}(X_k)$ estimates in (27) be continuous on E^k and absolutely bounded by some $B > 0$ for every $1 \leq \ell \leq q$, $1 \leq i \leq p$.

Let

$$\delta\left(\frac{\epsilon}{2\sum_{\ell=1}^{q}|a_\ell|}, k, B\right) > 0$$

be a constant such that: if Q_k^1 is satisfying

$$d_{L\rho_o}(Q_k^o, Q_k^1) < \delta\left(\frac{\epsilon}{2\sum_{\ell=1}^{q}|a_\ell|}, k, B\right)$$

then

$$\sum_{\ell=1}^{q}|a_\ell||m_{i\ell}^o - m_{i\ell}^1| < \frac{\epsilon}{2}$$

for some given $\epsilon > 0$. In this case, if for the same above given $\epsilon > 0$, and for some natural number n_a , there is some $\delta(\epsilon, n_a) > 0$:

$$\delta(\epsilon, n_a) \leq \delta\left(\frac{\epsilon}{2\sum_{\ell=1}^{q}|a_\ell|}, k, B\right)$$

that for every $n \geq n_a$ and Q_k^1 satisfying

$$d_{L,\rho_o}(Q_k^o, Q_k^1) < \delta(\epsilon, n_a)$$

there is some D_k^{o1} with Q_k^o, Q_k^1 marginals that satisfies the condition:

$$D_k^{o1}\{|\frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{q}a_\ell[[\mu_{i\ell}(X_{k,j\ell}) - m_{i\ell}^o] - [\mu_{i\ell}(Y_{k,j\ell}) - m_{i\ell}^1]]| \geq \frac{\epsilon}{2}\} \leq \frac{\epsilon}{p}$$

$$\forall\ 1 \leq i \leq p \qquad (39)$$

Then the estimate in (27) is Lèvy-Lèvy weakly robust by definition 4.

The boundness condition on the estimates that appeared in section 4 is included again in corollary 2. As mentioned before, this condition is a realistic property that protects the estimate values from wandering in the

space of unacceptable values.

We will now concentrate our attention on expression (39), hoping to obtain a comparatively small lower bound on the sample size n that satisfies the bound ϵ/p. That, of course, we expect for particular $\mu_{i\ell}(\cdot)$ choices.

We will point out again that the brackets

$$[[\mu_{i\ell}(X_{k,j\ell}) - m_{i\ell}^{o}] - [\mu_{i\ell}(Y_{k,j\ell}) - m_{i\ell}^{1}]] \tag{40}$$

in (39) are zero mean independent variables for every different $j\ell$ value.

A theorem expressed by Revész ([5], pg. 57) will be extremely useful here. We state the theorem below.

Theorem 1

Let $x_1, x_2, \ldots, x_n$ be independent, zero mean, not necessarily identically distributed variables. Then, the probability

$$P_n(\eta) = P_r\{\left|\frac{x_1 + x_2 + \ldots + x_n}{n}\right| \geq \eta\}$$

converges to zero exponentially for any $\eta > 0$, i.e., there is some $C > 0$ and some $0 < \nu < 1$ such that:

$$P_n(\eta) \leq C\nu^n$$

if and only if: For all $\eta > 0$ there exists a constant $\mathcal{E}_\eta > 0$ and some $t_\eta > 0$ such that:

$$\prod_{k=1}^{n} E\{e^{tx_k}\} \leq \mathcal{E}_\eta\, e^{|t|\eta n} \quad \text{whenever} \quad |t| \leq t_\eta$$

Also, the probability $P_n(\eta)$ cannot converge faster than exponentially and the constants C and ν that express the bound on the probability $P_n(\eta)$ are chosen as follows: $C = \mathcal{E}_\eta$; $\nu = e^{(\delta-\eta)\cdot t_\eta}$; where δ some value in the interval $(0,\eta)$.

We will apply the above theorem on condition (39) to design robust estimates that, in addition to satisfying the continuity and boundness properties expressed in corollary 2, they also have the strong characteristic of the fastest possible convergence to their asymptotic value, which through the absolute boundness is guaranteed to be stable inside stationary Q_k^o-Lèvy-contaminated distribution families, whenever the contamination is small enough.

Directly from theorem 1, from the fact that the expressions in (40), are independent and zero mean for different $j\ell$ values and from the observation that the sums

$$\sum_{\ell=1}^{q} a_\ell [[\mu_{i\ell}(X_{k,j\ell}) - m_{i\ell}^o] - [\mu_{i\ell}(Y_{k,j\ell}) - m_{i\ell}^1]] \tag{41}$$

are identically distributed for every j, we obtain that the left part of inequality (39) converges exponentially if and only if:

For all $\epsilon > 0$ there exists some constant $\mathcal{E}_{\epsilon/2} > 0$ and some $t_{\epsilon/2} > 0$ such that

$$\left[\prod_{1\le\ell\le q} E_{D_k^{o1}}\left\{e^{ta_\ell[[\mu_{i\ell}(X_k)-m_{i\ell}^o]-[\mu_{i\ell}(Y_k)-m_{i\ell}^1]]}\right\}\right]^n \le \mathcal{E}_{\epsilon/2}\, e^{|t|\frac{\epsilon}{2}n} \tag{42}$$

$$\forall\ |t| \le t_{\epsilon/2}$$

Due to theorem 1, the larger $t_{\epsilon/2}$ we can find the faster the convergence of the estimators to their asymptotic value. Also, we must emphasize here that we are seeking a $t_{\epsilon/2}$ that is common for all Q_k^1 that are members of the data distribution contaminated family.

As an additional observation on (42), we see that since its left part is equal to one for $t = 0$, $\mathcal{E}_{\epsilon/2}$ cannot be smaller than one. Seeking the smallest possible $\mathcal{E}_{\epsilon/2}$ we may as well pick $\mathcal{E}_{\epsilon/2} = 1$.

In this case, condition (42) can take the form:

$$\sum_{\ell=1}^{q} \ell n E_{D_k^{o1}} \{ e^{t a_\ell [[\mu_{i\ell}(X_k) - m_{i\ell}^o] - [\mu_{i\ell}(Y_k) - m_{i\ell}^1]]} \} \leq |t| \frac{\epsilon}{2}$$

$$\forall\ |t| \leq t_{\epsilon/2} \qquad (43)$$

where $t_{\epsilon/2}$ some positive value and D_k^{o1} some 2k-dimensional distribution with Q_k^o, Q_k^1 marginals. Each of the logarithmic expressions in (43) is a convex U function of t with minimum at $t = 0$ and minimum value equal to zero. This is true done to the fact that each of these logarithmic functions has positive second derivative for every t and first derivative at $t = 0$ that is equal to zero. The above observations are true for every Q_k^o, Q_k^1, D_k^{o1} choice. Also, the sum of the logarithmic functions in (43) is also convex U with minimum that is equal to zero and happens at $t = 0$.

Due to the above observations, there is always some $t_{\epsilon/2}$ (for given $\epsilon > 0$) that satisfies (43). The analysis should be concentrated now to designing the constants a_ℓ and the functions $\mu_{i\ell}(\cdot)$ in a way that will make the common for all distributions in the contaminated family $t_{\epsilon/2}$ as large as possible.

Before we express some thoughts on that we will state the following theorem:

Theorem 2

Let the set of estimates in (27) be continuous E^k and absolutely bounded by some $B > 0$ for every $1 \leq \ell \leq q$; $1 \leq i \leq p$. Then, if for some $\delta > 0$ and such that it is nonlarger than the

$$\delta\left(\frac{\epsilon}{2\sum_{\ell=1}^{q} |a_\ell|}, k, B\right)$$

included in corollary 2, a common $t_{\epsilon/2} > 0$ can be found for all the members of the Q_k^o-contaminated family characterized by:

$$D_{L\rho_o}(Q_o^k, Q_k) \leq \delta \tag{44}$$

where $t_{\epsilon/2}$ is such that it satisfies the expression

$$\sum_{\ell=1}^{q} \ell n \; E_{Q_k^o}\{e^{ta_\ell[\mu_{i\ell}(X_k)-m_{i\ell}^o]}\} E_{Q_k^1}\{e^{-ta_\ell[\mu_{i\ell}(X_k)-m_{i\ell}^1]}\} \leq |t|\frac{\epsilon}{2} \tag{45}$$

$$\forall\ |t| \leq t_{\epsilon/2};\ \forall\ Q_k^1 \text{ satisfying (44)}$$

the estimate in (27) is weakly robust and it converges to its asymptotic value with rate expressed by the bound:

$$e^{n(\zeta-\frac{\epsilon}{2})t_{\epsilon/2}};\ \zeta \in (0, \tfrac{\epsilon}{2}) \tag{46}$$

The absolute boundness of $\mu_{i\ell}(\cdot)$ guarantees stability of the $\mu_{i\ell}(\cdot)$ asymptotic value inside the contaminated family expressed by (44).

Also, for performance equal to ϵ (where $\epsilon > 0$) it is sufficient that we choose then number of samples equal to:

$$n_a = \frac{\ell n \frac{\epsilon}{p}}{(\zeta - \frac{\epsilon}{2})t_{\epsilon/2}} \tag{47}$$

ζ is some value in $(0,\epsilon/2)$ and it is the same in both expressions (46) and (47).

In theorem 2 we picked $D_k^{o1}(X,Y) = Q_k^o(X)\cdot Q_k^1(Y)$. We will now concentrate on expression (45) to draw some useful conclusions. Indeed we can separate the expression (45) into two parts: One including the behavior of the estimate at the central distribution Q_k^o and one describing the same behavior at the distributions included in the contaminated family in (44).

Through the separation we just mentioned we can write (45) in the following way:

$$\sum_{\ell=1}^{q} \ln E_{Q_k^o}\{e^{ta_\ell[\mu_{i\ell}(X_k)-m_{i\ell}^o]}\} + \sum_{\ell=1}^{q} \ln E_{Q_k^1}\{e^{-ta_\ell[\mu_{i\ell}(X_n)-m_{i\ell}^1]}\} \leq |t|\frac{\epsilon}{2}$$

$$\forall\ |t| \leq t_{\epsilon/2} \tag{48}$$

Both sums on the left part of (48) are convex U with minimum equal to zero and assumed at $t = 0$. The less sharp both of these sums are as functions of t, the larger $t_{\epsilon/2}$ will be for given $\epsilon > 0$. Since Q_k^o is a well-defined distribution, the set $\{a_\ell, \mu_{i\ell}(\cdot);\ 1 \leq \ell \leq q,\ 1 \leq i \leq p\}$ can be designed to make the sum

$$\sum_{\ell=1}^{q} \ln E_{Q_k^o}\{e^{ta_\ell[\mu_{i\ell}(X_k) - m_{i\ell}^o]}\} \tag{49}$$

as less sharp at $t = 0$ as possible.

We will call an estimate $\{a_\ell, \mu_{i\ell}(X_k);\ 1 \leq \ell \leq q;\ 1 \leq i \leq p\}$ in (27) a fast estimate at Q_k^o if it makes the expression in (49) a slowly increasing function of t around $t = 0$. The continuity of the functions $\mu_{i\ell}(\cdot)$ on E^k and their boundness, guarantees closeness of the moment generating functions

$$E_{Q_k^1}\{e^{-ta_\ell[\mu_{i\ell}(X_k)-m_{i\ell}^o]}\}$$

to the central moment generating function $E_{Q_k^o}\{e^{-ta_\ell[\mu_{i\ell}(X_k)-m_{i\ell}^o]}\}$ for Q_k^1 belonging to the contaminated family (44). Therefore, for fast exponential convergence, then, it is sufficient to design the estimator in (27) to make the logarithmic expression of the moment generating functions at Q_k^o (expression (49)) as slowly increasing with t around $t = 0$ as possible.

As conclusion, we finally express the following theorem:

Theorem 3

If the set of estimates in (27) is continuous as a function on E^k, absolutely bounded for every $1 \leq \ell \leq q$, $1 \leq i \leq p$ and it is a fast estimate

at Q_k^o, then it is also weakly robust at Q_k^o, and it converges fast and exponentially to its stable mean.

The advantage of the method presented in this section, in comparison to Hampel's method is that the design of the weakly robust estimates reduces to real function properties and to moment generating functions at the central distribution Q_k^o. Furthermore, this method has the tremendous advantage of treating finite data samples (through the smallest n_a that achieves the required performance) and not just asymptotic situations as in [1]. Finally, the Lèvy-performance criterion is stronger than the small-variance criterion treated by Huber [1].

Finding the smallest $t_{\epsilon/2}$ in theorem 2 that will satisfy all members of the Q_k^o-contaminated family is a task, samples of which will be shown in the following section.

Concluding this section, we will mention that theorem 1 has also been applied in [19] to find confidence intervals for Bayes error estimation which were subsequently used to determine the optimal degree of quantization.

6. FURTHER DISCUSSION ON SECTIONS 3, 4 AND 5

In this section we will present some discussion on the application of the theory that appears in the previous sections. Our discussion will be mostly oriented toward expressing methods and suggesting ideas for the actual design of the robust estimators under consideration. The possibilities for such designs are many and lengthy and they will appear in detail in future work (under preparation).

1. We will first start with the Vasershtein robust estimators that are analyzed in section 3. According to lemma 2, such an estimator (where the underlying distortion measure is the square error) must be linear. That is:

$$\hat{S}_n(X_n) = \{\hat{s}_{in}(X_n);\ i = 1,\ldots,p\};\ \hat{s}_{in}(X_n) = \sum_{k=1}^{n} a_{ni}(k)x_k \qquad (50)$$

Also, a sufficient condition for robustness in this case is that each of the sequences

$$\{\sum_{k=1}^{n} |a_{ni}(k)|\}\ ;\ i = 1,\ldots,p$$

converges to some finite value. Any

$$\{\sum_{k=1}^{n} |a_{ni}(k)|\}$$

that is geometric with multiplying factor ω smaller than one belongs to the above group. Also, the experimental mean is obviously Vasershtein robust.

In general, one will look for linear estimates that converge at the central well-known distribution to the desired value and whose coefficients satisfy the absolute convergence property mentioned above. For example, let the central distribution Q^o be m-dependent. Then accumulate the data in m-groups and form the estimate

$$\hat{s}_{i,nm}(X_{nm}) = \frac{1}{n}\sum_{k=1}^{nm} a^1_{nm,i}(k)x_k \qquad (51)$$

where the previous coefficients $a_{ni}(k)$ in (50) are related to the $a^1_{nm,i}(k)$'s in (51) by

$$a_{nm,i}(k) = \frac{1}{n}\, a^1_{nm,i}(k) \qquad (52)$$

Due to the m-dependence of Q^o, the sums

$$\sum_{k=1}^{m} a^1_{nm,i}(k)x_k, \ldots, \sum_{k=jm}^{(j+1)m} a^1_{nm,i}(k)x_k, \ldots$$

are independent from each other and according to theorem 1, the estimate in (51) converges exponentially and in probability to

$$m_o \cdot \lim_{n\to\infty} n^{-1} \sum_{k=1}^{nm} a_{nm,i}^{1}(k) = m_o \cdot \lim_{n\to\infty} \sum_{k=1}^{n} a_{ni}(k) \tag{53}$$

where m_o the one-dimensional mean determined by Q^o .

If we pick all the coefficients positive and design the sum $\sum_{k=1}^{n} a_{n_i}(k)$ so that it converges to the desired value at Q^o , then at the same time the robustness sufficient condition is also satisfied and the estimate in (51) is stable inside the Vasershtein contaminated family.

2. For the Lèvy-Vasershtein robustness (as found in section 4), one should seek for continuous as real functions and continuous at the central distribution estimates that are also absolutely bounded. This last property is the only additional to the ones required by the robust in the Hampel sense estimates.

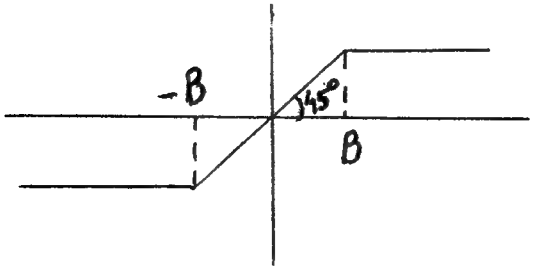
Of course, all estimates that are not Hampel robust (such as the experimental mean), are not robust in the present sense either.

Also, the L-estimators

$$\hat{s}_{in}(X_n) = \sum_{j=1}^{n} b_j x_{(j)} \tag{54}$$

where $X_n = \{x_i\}$, $x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$, the ordered components of X_n, are not in general absolutely bounded, therefore not Lèvy-Vasershtein robust in general.

On the contrary, a truncated version of the estimates in (54) can be Lèvy-Vasershtein robust with proper choice of the coefficients b_j . By truncated version of (54) we mean that we substitute each $x_{(j)}$ in it by its output when passed through the following nonlinearity:



Finally, using the approach of the M-estimators, we can find the estimate $\hat{s}_{in}(X_n)$ as the truncated zeros of a sum [18]:

$$\sum_{j=1}^{n} \psi_i(x_j - \hat{s}_{in}(X_n)) \tag{55}$$

where $\psi_i(\cdot)$ a smooth enough function. The zeros of the expression in (55) will be passed through a nonlinearity as the one mentioned above, or any other that cuts off all estimate values that are absolutely higher than B.

3. The fast robustness introduced in section 5 is important and deserves special attention. Indeed, the method introduced there can be considered small sample since the effort is toward designing estimates that, besides being robust, converge to their stable (for small deviations of the data distribution from the central one) mean fast.

The measure of performance is Lèvy distance. In other words, if a specific contaminated family of data distributions is given, the stability of the estimate in it is measured through the longest Lèvy distance of the estimate distribution at some Q from the estimate distribution at Q^o, where Q^o is the central well-known data distribution and Q moves inside the contaminated family.

As explained in section 5, if p scalar estimates are calculated from the same data and the performance required is a fixed $\epsilon > 0$, then the minimum number of data n_a that will satisfy this performance is given by

$$n_a = \frac{\ell n \frac{\epsilon}{p}}{(\zeta - \frac{\epsilon}{2}) t_{\epsilon/4}} \tag{56}$$

where ζ is some arbitrary constant in $(0, \epsilon/2)$ and $t_{\epsilon/4}$ is for given estimates (of (27) form and k-dependent group data) the smallest $t_{Q_k^1, \epsilon/4}$ such that

$$\sum_{\ell=1}^{q} \ln E_{Q_k^1}\{e^{-ta_\ell[\mu_{i\ell}(X_k)-m_{i\ell}^1]}\} \leq |t|\frac{\varepsilon}{4} \; ; \; \forall \; |t| \leq t_{Q_k^1,\varepsilon/4} \tag{57}$$

where Q_k^1 moves in the considered data contaminated family and the smallest $t_{Q_k^1,\varepsilon/4}$ is taken among these Q_k^1 members.

To achieve this minimum $t_{\varepsilon/4}$ as large as possible, we go backward. Specifically, we are looking for this estimate set $\{a_\ell; \mu_{i\ell}(X_k); 1 \leq \ell \leq q; 1 \leq i \leq p\}$ that achieves the largest $t_{Q_k^1,\varepsilon/4}$ for the worst Q_k^1 choice inside the specific given contaminated family. This worst case corresponds to

$$\sum_{\ell=1}^{q} \ln E_{a_k}\{e^{-ta_\ell[\mu_{i\ell}(X_k)-m_{i\ell}^1]}\}$$

function with the fastest increase for $|t|$ increasing.

Specific continuous and absolutely bounded functions $\mu_{i\ell}(X_k)$ can be chosen with some of their characteristics left flexible. The design of these characteristics becomes a maximum problem (find largest $t_{Q_k^1,t_{\varepsilon/4}}$ for the logarithmic moment generating expression the sharpest in the family) and the methods for solving it are similar to the ones applied in [1] and [19].

REFERENCES

1. P.J. Huber (1964) "Robust estimation of a location parameter," Ann. Math Statist. 36: 1753-1758.

2. V. Strassen (1965) "The existence of probability measures with given marginals," Ann. Math. Statist. 36: 423-439.

3. P.J. Huber (1967) "The behavior of maximum likelihood estimates under nonstandard conditions," Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1: 221-233.

4. R.M. Dudley (1968) "Distances of probability measures and random variables," Ann. Math. Statist. 39: 1563-1572.

5. Pál Revész (1968) The Laws of Large Numbers in Probability and Mathematical Statistics. A series of monographs and textbooks edited by Z.W. Birnbaum and E. Lukacs, Academic Press, New York.

6. L.N. Vasershtein (1969) "Markov processes on a countable product space, describing large systems of automata," Problemy Peredachi Informatsii, 5(3): 64-73 (in Russian).

7. P.J. Huber (1970) "Studentizing Robust Estimates," in Nonparametric Techniques in Statistical Inference, M.L. Puri, ed., Cambridge Univ. Press: 453-463.

8. R.L. Dobrushin (1970) "Prescribing a system of random variables by conditional distributions," Theory of Prob. and Appl. XV(3):458-486.

9. L.N. Vasershtein and A.M. Leonfovich (1970) "On invarient measures of certain Markov operators, describing a random medium," Problemy Peredachi Informatsii 6(1): 71-80 (in Russian).

10. F.R. Hampel (1971) "A general qualitative definition of robustness," Ann. Math. Statist. 42: 1887-1896.

11. R.D. Martin and S.C. Schwartz (1971) "Robust detection of a known signal in nearly Gaussian noise," IEEE Trans. Inf. Th. IT-17: 50-56.

12. P.J. Huber (1972) "Robust statistics. A review," Ann. Math. Statist. 43: 1041-1067.

13. R.D. Martin and C.J. Masreliez (1975) "Robust estimation via stochastic approximation," IEEE Trans. Inf. Th. IT-21: 257-262.

14. R.M. Gray, D. Neuhoff and P.C. Shields (1975) "A generalization of Orstein's distance with applications to information theory," Ann. Prob. 3(2): 315-324.

15. D.L. Neuhoff, R.M. Gray and L.D. Davisson (1975) "Fixed rate universal block source coding with a fidelity criterion," IEEE Trans. Inf. Th. IT-21(5): 511-523.

16. P. Papantoni-Kazakos (1975) "Small sample efficiencies of rank tests," IEEE Trans. Inf. Th. IT-21: 150-156.

17. P. Papantoni-Kazakos (1976) "Some distance measures and their use in feature selection," Rice University Technical Report #7611, Electrical Engineering Department, Rice University, Houston, Texas.

18. P. Papantoni-Kazakos (1977) "Robustness in parameter estimation," IEEE Trans. Inf. Th., accepted for publication, March 1977.

19. D. Kazakos (1977) "Quantization complexity and training data sample size in detection problems," submitted for publication.

20. S.A. Kassam and J.B. Thomas (1975) "A class of nonparametric detectors of dependent input data," IEEE Trans. Inf. Th. IT-21: 431-437.

21. S.A. Kassam and J.B. Thomas (1976) "Asymptotically robust detection of a known signal in contaminated non-Gaussian noise," IEEE Trans. Inf. Th. IT-22: 22-26.

APPENDIX

Proof of Lemma 3

In the distance $d_{Vt}(D^o(\hat{S}_n),D(\hat{S}))$ in (25), the terms

$$(\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)})^2$$

and $(m^o_{in} - m_{in})^2$ appear. We develop upper bounds on each of them. Indeed, we obtain:

$$(\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)})^2 < |\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)}|\,|\sqrt{\sigma^o_{in}(0)} + \sqrt{\sigma_{in}(0)}\,|$$

$$= |\sigma^o_{in}(0) - \sigma_{in}(0)| = \left|\int \hat{s}^2_{in}(X_n)Q_o(dX_n) - \int \hat{s}^2_{in}(X_n)Q(dX_n)\right.$$

$$\left. - [(m^o_{in})^2 - (m_{in})^2]\right|$$

$$\leq \left|\int \hat{s}^2_{in}(X_n)Q_o(dX_n) - \int \hat{s}^2_{in}(X_n)Q(dX_n)\right| + |(m^o_{in})^2 - (m_{in})^2| \quad (A.1)$$

If the estimate $\hat{s}_{in}(X_n)$ is absolutely bounded from above by some $A_i > 0$, then we obtain from (A.1):

$$(\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)})^2 < \int |\hat{s}_{in}(X_n) + \hat{s}_{in}(Y_n)|\,|\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| \cdot$$

$$\cdot D(dX_n,dY_n) + 2A_i\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)|D(dX_n,dY_n)$$

$$\leq 4A_i\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)|D(dX_n,dY_n) \quad (A.2)$$

where $D(X_n,Y_n)$ some distribution with $Q^o(X_n)$, $Q(Y_n)$ marginals. From (A.2) we also obtain, for the same $D(X_n,Y_n)$:

$$(\sqrt{\sigma^o_{in}(0)} - \sqrt{\sigma_{in}(0)})^2 + (m^o_{in} - m_{in})^2$$

$$< 6A_i\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)|D(dX_n,dY_n)$$

For robustness from $d_L(Q_o,Q)$ to $d_V(D^o(\hat{S}_n),D(\hat{S}_n))$ we require:

That given $\epsilon > 0$, there is some $\delta(\epsilon) > 0$ such that

If $d_L(Q_o,Q) < \delta(\epsilon) \Rightarrow d_V(D^o(\hat{S}_n),D(\hat{S}_n)) < \epsilon$; $\forall$ n .

Or:

Given $\epsilon > 0$, there is some $\delta(\epsilon) > 0$:

For $d_L(Q_o,Q) < \delta(\epsilon)$

$$\Rightarrow \max_{1\leq i\leq p} (6A_i \int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n,dY_n)) < \epsilon \ ; \ \forall\, n$$

where $D(X_n,Y_n)$ has respective marginals $Q_n^o(X_n)$ and $Q_n(Y_n)$. If given $\epsilon > 0$, there is for every $1 \leq i \leq p$ some $\delta_i(\epsilon) > 0$:

$$\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n,dY_n) < \frac{\epsilon}{6A_i} \ ; \ \forall\, n$$

Then

$$\delta(\epsilon) = \min_{1\leq i\leq p} \delta_i(\epsilon)$$

If $d_L(Q_n^o,Q_n) < \delta_{in}(\epsilon)$, then there is some $\forall\, n$ $D(X_n,Y_n)$ with $Q_n^o(X_n),Q(Y_n)$ marginals (Strassen [2], Th. 11) such that

$$D(X_n,Y_n : \rho(X_n,Y_n) \geq \delta_{in}(\epsilon)$$

Pick this $D(X_n,Y_n)$ and write:

$$\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n,dY_n)$$

$$= \int_{X_n,Y_n : \rho(X_n,Y_n) < \delta_{in}(\epsilon)} |\hat{s}_{in}(Y_n) - \hat{s}_{in}(X_n)| D(dX_n,dY_n) + \int_{X_n,Y_n : \rho(X_n,Y_n) \geq \delta_{in}(\epsilon)} |\hat{s}_{in}(Y_n) - \hat{s}_{in}(X_n)| D(dX_n,dY_n)$$

$$\leq 2A_i \delta_{in}(\epsilon) + \int_{X_n,Y_n : \rho(X_n,Y_n) < \delta_{in}(\epsilon)} |\hat{s}_{in}(Y_n) - \hat{s}_{in}(X_n)| D(dX_n,dY_n)$$

since $|\hat{s}_{in}(X_n)| \leq A_i \ ; \ \forall\, n$

Let $\hat{s}_{in}(X_n)$ be continuous function everywhere on E^n . Then, given $\epsilon_i > 0$ and X_n , there is some $\delta(\epsilon_i,X_n) > 0$ such that

$$\forall\, Y_n : \rho(X_n,Y_n) < \delta(\epsilon_i,X_n) \Rightarrow |\hat{s}_{in}(Y_n) - \hat{s}_{in}(X_n)| < \epsilon_i$$

Choose a sequence $\{c_j\}$ of positive, monotonically decreasing toward zero components and define

$$A_j = \{X_n : \delta(\epsilon_j, X_n) > c_j\}$$

Then

$$A_j \subseteq A_{j+1} \rightarrow E^n \; ; \; UA_j = E^n$$

Define

$$\underset{\delta(\epsilon_i, X_n)}{U} (X_n) = \{Y_n : \rho(X_n, Y_n) < \delta(\epsilon_i, X_n)\}$$

$$B_j = U\left\{ \underset{\delta(\epsilon_i, X_n)}{U} (X_n) : X_n \in A_j \right\}$$

Then $B_j \subseteq B_{j+1}$, $UB_j = E^n$

Now, given $\eta_i > 0$, there is some $k_i(n, \eta_i)$ such that

$$Q_o\left(\underset{1 \leq j \leq k_i(n,\eta_i)}{UB_j} \right) > 1 - \eta_i$$

Denote: $\mathcal{E}_{in} = \underset{1 \leq j \leq k_i(n,\eta_i)}{UB_j}$

Pick

$$\epsilon_i = \frac{\epsilon}{18A_i} \; , \; \eta_i = \frac{\epsilon}{36A_1^2}$$

$$\delta_{in}(\epsilon) = \min\left\{ c_{k_1}\left(n, \frac{\epsilon}{36A_1^2}\right), \frac{\epsilon}{36A_1^2} \right\}$$

Then

$$\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n)$$

$$\leq 2A_i \delta_{in}(\epsilon) + \int_{\substack{X_n \in \mathcal{E}_{in} \\ Y_n : \rho(X_n,Y_n) < \delta_{in}(\epsilon)}} |\hat{s}_{in}(Y_n) - \hat{s}_{in}(X_n)| D(dX_n, dY_n)$$

$$+ \int_{\substack{X_n \in \mathcal{E}_{in}^c \\ Y_n : \rho(X_n,Y_n) < \delta_{in}(\epsilon)}} |\hat{s}_{in}(Y_n) - \hat{s}_{in}(X_n)| D(dX_n, dY_n)$$

$$\leq 2A_i \delta_{in}(\epsilon) + 2A_i \eta_i + \int_{\substack{X_n \in \mathcal{E}_{in} \\ Y_n : \rho(X_n,Y_n) < \delta_{in}(\epsilon)}} |\hat{s}_{in}(Y_n) - \hat{s}_{in}(X_n)| D(dX_n, dY_n)$$

$$\leq 2A_i \delta_{in}(\epsilon) + 2A_i \eta_i + \epsilon_i$$

$$\leq 2A_i \frac{\epsilon}{36A_i^2} + 2A_i \frac{\epsilon}{36A_i^2} + \frac{\epsilon}{18A_i} = \frac{\epsilon}{6A_i}$$

For $n \leq n_o$, we can pick

$$\Rightarrow \delta_{in_o}(\epsilon) = \min\left[\frac{\epsilon}{36A_i^2}, \min_{n \leq n_o} C_{k_i}\left(n, \frac{\epsilon}{36A_i^2}\right)\right]$$

to satisfy

$$d_L(Q_{on}, Q_n) < \delta_{in}(\epsilon) = \delta_{in_o}(\epsilon)$$

$$\Rightarrow \int |\hat{s}_{in}(X_{in}) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n) < \frac{\epsilon}{6A_i} \; ; \; \forall \; n \leq n_o$$

Proof of Lemma 4

To take care of $n > n_o$, let $d_L(Q^o, Q) < \delta_n(\epsilon)$ for $n > n_o$, where $\epsilon_i > 0$ is given. Pick again $D(X_n, Y_n)$ with Q_{on}, Q_n marginals such that

$$D(X_n, Y_n : \rho(X_n, Y_n) \geq \delta_n(\epsilon_i) \leq \delta_n(\epsilon_i)$$

and write

$$\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n) \leq 2A_i \delta_n(\epsilon_i) +$$

$$+ \int_{X_n, Y_n : \rho(X_n, Y_n) < \delta_n(\epsilon_i)} |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n) \; ; \; n > n_o$$

For arbitrary $Q^o(\cdot)$ there is nothing that can be done at this point. So, let $Q^o(\cdot)$ be m dependent. That is, if $x_1, x_2, \ldots, x_n$ is distributed according to $Q_o(\cdot)$, then every x_i depends on only m preceding and following data.

Then, let X_n consist of X_{ki}; i = 1,2,... vectors of consecutive data that have m data gaps between X_{ki} and $X_{k,i+1}$. Then, X_{ki}, X_{kj}; $i \neq j$ are independent by $Q^o(\cdot)$. Let $n = kn_1$ and define the following experimental distribution:

$$n^k_{X_n}(\omega_1, \ldots, \omega_k) = \frac{1}{n_1} \{\# \text{ of } X_k\text{'s with } x_1 < \omega_1, x_2 < \omega_2, \ldots, x_k < \omega_k\}$$

The typical sequences X_n are such that $n^k_{X_n}(\omega_1, \ldots, \omega_k)$ approaches $Q^o_k(X_k)$ for n large enough.

Or, given $\eta_i > 0$, there is some n_o and some $\delta(\eta_i) > 0$ such that

$$\forall\, n > n_o \Rightarrow Q^o\{d_{L_\rho}(Q^o_k, n^k_{X_n}) \geq \delta(\eta_i)\} \leq \eta_i$$

Let

$$\mathcal{E}_{ni} = \{X_n : d_{L_\rho}(Q^o_k, n^k_{X_n}) < \delta(\eta_i)\}$$

Then

$$Q^o(\mathcal{E}_{ni}) > 1 - \eta_i \; ; \; \forall\, n > n_o$$

Going back on page 18 we have now

$$\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n) \leq 2A_i \delta_n(\epsilon_i)$$

$$+ \int_{\substack{X_n \in \mathcal{E}_{ni} \\ Y_n : \rho(X_n, Y_n) < \delta_n(\epsilon_i)}} |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n) + \int_{\substack{X_n \in \mathcal{E}_{ni}^c \\ Y_n : \rho(X_n, Y_n) < \delta_n(\epsilon_i)}} |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n)$$

$$\leq 2A_i \delta_n(\epsilon_i) + 2A_i \eta_i + \int_{\substack{X_n \in \mathcal{E}_{ni} \\ Y_n : \rho(X_n, Y_n) < \delta_n(\epsilon_i)}} |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n)$$

Let

$$\rho(X_n, Y_n) < \delta_n(\epsilon_i) \Rightarrow d_{L_\rho}(n^k_{X_n}, n^k_{Y_n}) < \zeta_\rho(\delta(\epsilon_i)) \; ; \; \forall n > n_o$$

where ζ_ρ depends on the measure $\rho(\cdot,\cdot)$. Then,

$$\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n)$$

$$\leq 2A_i \delta_n(\epsilon_i) + 2A_i \eta_i + \int_{\substack{X_n : d_{L_\rho}(Q^o_k, n^k_{X_n}) < \delta(\eta_i) \\ Y_n : d_{L_\rho}(n^k_{X_n}, n^k_{Y_n}) < \zeta_\rho(\delta(\epsilon_i))}} |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n)$$

$d_{L_\rho}(\cdot,\cdot)$ satisfying the triangle inequality and, being symmetric, we have:

$$d_{L_\rho}(Q^o_k, n^k_{Y_n}) \leq d_{L_\rho}(Q^o_k, n^k_{X_n}) + d_{L_\rho}(n^k_{X_n}, n^k_{Y_n})$$

So, from above we have

$$\int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n) \leq 2A_i \cdot \delta(\epsilon_i) + 2A_i \eta_i$$

$$+ \int_{\substack{X_n, Y_n : d_{L_\rho}(Q^o_k, n^k_{X_n}) < \delta(\eta_i) \\ d_{L_\rho}(Q^o_k, n^k_{Y_n}) < \delta(\eta_i) + \zeta_\rho(\delta(\epsilon_i))}} |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n)$$

Let $\hat{s}_{in}(X_n)$ be continuous at Q_{ok}. That is, given $\epsilon_i/3 > 0$, there is some $\mu(\epsilon_i/3) > 0$ such that

$$\forall X_n : d_{L_\rho}(n^k_{X_n}, Q^o_k) < \mu(\epsilon_i/3)$$

$$\Rightarrow |\hat{s}_{in}(X_n) - s_i(Q^o_k)| < \epsilon_i/3$$

where $s_i(Q^o_k)$ the value that $\hat{s}_{in}(X_n)$ converges for $n \to \infty$.

Choose:

$$\eta_i = \frac{\epsilon_i}{6A_i} = \frac{\epsilon}{36A_i^2} \; ; \; \epsilon_i = \frac{\epsilon}{6A_i}$$

$$\delta(\eta_i) = \min(\delta(\eta_i = \frac{\epsilon}{36A_i^2}), \tfrac{1}{2}\mu(\frac{\epsilon}{18A_i}))$$

$$\delta(\epsilon_i) = \min(\zeta_\rho^{-1}(\tfrac{1}{2}\mu[\frac{\epsilon}{18A_i}]), \frac{\epsilon}{36A_i^2})$$

Then

$$2A_i\delta(\epsilon_i) + 2A_i\eta_i + \int |\hat{s}_{in}(X_n) - \hat{s}_{in}(Y_n)| D(dX_n, dY_n)$$

$$X_n, Y_n : d_{L_\rho}(Q^o_k, n^k_{X_n}) < \delta(\eta_i)$$

$$d_{L_\rho}(Q^o_k, n^k_{Y_n}) < \delta(\eta_i) + \zeta_\rho(\delta(\epsilon_i))$$

$$\leq \frac{\epsilon}{18A_i} + \frac{\epsilon}{18A_i} + \frac{\epsilon}{18A_i} = \frac{\epsilon}{6A_i}$$

Finally, pick for every i :

$$\delta(\epsilon) = \min_i \delta(\epsilon_i) =$$

$$= \min_i \min\{\frac{\epsilon}{48A_1^2}, \min_{n \leq n_o} c_{k_i}(n, \frac{\epsilon}{48A_i^2}), \zeta_\rho^{-1}(\tfrac{1}{2}\mu[\frac{\epsilon}{24A_i}])\}$$

and then $d_V(D^o(\hat{S}_n), D(\hat{S}_n)) < \epsilon$,

for $d_{L_\rho}(Q^o, Q) < \delta(\epsilon)$.